PR #50595 完整报告

vllm-project/vllm

[Bugfix][Structured Output] Mask request stop tokens in xgrammar until grammar terminates

合并时间：2026-08-14 02:07

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/50595

执行摘要

- 一句话：清理 Harmony stop token 补丁并更新测试
- 推荐动作：建议结合 #49227 一起阅读，理解结构化输出后端（XgrammarBackend）的 stop token 掩码抽象，以及测试如何镜像生产路径。本 PR 小而清晰，是观察 vLLM 如何在后端能力落地后清理解析器 workaround 的好样例。

功能与动机

PR body 明确说明：Following #49227's merge, remove the patch used in HarmonyParser and update test to mirror production path。源码注释原本留有 TODO: Remove <|call|> once #50595 lands.，本 PR 正是落地这一清理；同时 bbrowning 在 issue 评论中确认该变更修复了 #51693 的 500 错误。

实现拆解

1. 移除 HarmonyParser 中的 stop token 补丁：在 vllm/parser/harmony.py 中，_END_TAG 从包含 <|end|>、<|call|> 和空字符串的三元素列表收敛为仅包含 <|end|> 和空字符串，并更新注释说明原因（xgrammar 在约束下不允许 stop token，<|return|> 用空字符串表示）。这是 #49227 合并后的计划内清理，原注释留有 TODO: Remove <|call|> once #50595 lands。
2. 测试镜像生产路径：tests/parser/test_harmony.py 删除对 xgrammar.testing._is_grammar_accept_string 的依赖，改由 XgrammarBackend.compile_grammar 搭配 StructuredOutputOptions.STRUCTURAL_TAG 编译 grammar，并传入 stop_token_ids，最后用 grammar.validate_tokens 对 token 序列做校验。
3. 新增测试夹具：gpt_oss_stop_token_ids 从 GenerationConfig 读取 EOS token 集合，xgrammar_backend 构造 VllmConfig 并实例化 XgrammarBackend；两者被注入 test_adjust_request 与 _assert_structured_outputs_admission。
4. 验证：本地执行 pytest tests/parser/test_harmony.py 通过（64 passed, 38 warnings），bbrowning 拉取分支确认修复 #51693 场景。

关键文件：

- vllm/parser/harmony.py（模块 解析器；类别 source；类型 core-logic；符号 _END_TAG）：核心源码变更：移除 HarmonyParser 为规避旧 xgrammar 行为而加入的 <|call|> 补丁，

_END_TAG 收敛为仅包含 <|end|> 与空字符串，是本次 bugfix 的关键。

- tests/parser/test_harmony.py（模块 解析器；类别 test；类型 test-coverage；符号 gpt_oss_stop_token_ids, xgrammar_backend）：测试从字符串级 grammar 接受性检查改为通过真实 XgrammarBackend 编译并传入 stop_token_ids，新增 gpt_oss_stop_token_ids 与 xgrammar_backend fixtures，确保移除补丁后行为正确且覆盖 token 级校验。

关键符号：gpt_oss_stop_token_ids, xgrammar_backend, _assert_structured_outputs_admission, test_adjust_request

关键源码片段

vllm/parser/harmony.py

核心源码变更：移除 HarmonyParser 为规避旧 xgrammar 行为而加入的 <|call|> 补丁，_END_TAG 收敛为仅包含 <|end|> 与空字符串，是本次 bugfix 的关键。

```
# Harmony 的 stop tokens 是 <|return|>、<|call|>、<|endoftext|>。
# <|return|> 是默认 stop token，xgrammar 在约束下不允许 stop token，
# 否则会导致错误或无限生成，因此这里用空字符串 '' 表示。
# <|endoftext|> 不会被 StreamableParser 视为消息结束，所以不放入 end tag。
# 在 #49227 落地前需要把 <|call|> 一并放入 end tag 绕过限制；
# 现在 stop tokens 会在 grammar 终止前被掩码，<|call|> 不再需要特殊处理。
_END_TAG = ['<|end|>', '']
_FINAL_BEGIN = '<|channel|>final{constrain}<|message|>'
_TOOL_CALL_CHANNELS = [
    '<|channel|>commentary',
    '<|channel|>analysis',
    '<|channel|>final',
]
```

tests/parser/test_harmony.py

测试从字符串级 grammar 接受性检查改为通过真实 XgrammarBackend 编译并传入 stop_token_ids，新增 gpt_oss_stop_token_ids 与 xgrammar_backend fixtures，确保移除补丁后行为正确且覆盖 token 级校验。

```
# 直接从模型 GenerationConfig 读取 EOS token 作为 stop_token_ids。
# 这组 token 会传给 xgrammar 编译 grammar，用于在 grammar 终止前屏蔽请求级 stop tokens。
@pytest.fixture(scope='module')
def gpt_oss_stop_token_ids() -> set[int]:
    eos_token_id = GenerationConfig.from_pretrained(REASONING_MODEL_NAME).eos_token_id
    if isinstance(eos_token_id, int):
        return {eos_token_id}
    return set(eos_token_id)


# 构造与生产一致的 XgrammarBackend，而不是直接用 Grammar.from_structural_tag。
# 这样测试能覆盖 stop_token_ids 的掩码行为，避免与生产路径产生偏差。
@pytest.fixture(scope='module')
```

```
def xgrammar_backend(gpt_oss_tokenizer) -> XgrammarBackend:
    vllm_config = VllmConfig(
        structured_outputs_config=StructuredOutputsConfig(backend='xgrammar')
    )
    return XgrammarBackend(
        vllm_config,
        tokenizer=gpt_oss_tokenizer,
        vocab_size=len(gpt_oss_tokenizer),
    )


@classmethod
def _assert_structured_outputs_admission(
    cls,
    adjusted_request: ChatCompletionRequest | ResponsesRequest,
    expected_admission: Sequence[str],
    xgrammar_backend: XgrammarBackend,
    stop_token_ids: set[int],
) -> None:
    structured_outputs = adjusted_request.structured_outputs
    assert structured_outputs is not None
    assert structured_outputs.structural_tag is not None
    assert structured_outputs.all_non_structural_tag_constraints_none()

    # 通过真实后端编译 grammar，stop_token_ids 在生产中同样由调用方传入，
    # 测试与推理路径共用同一入口。
    grammar = xgrammar_backend.compile_grammar(
        StructuredOutputOptions.STRUCTURAL_TAG,
        structured_outputs.structural_tag,
        stop_token_ids=stop_token_ids,
    )
    expected_admission_set = set(expected_admission)

    for sample_name in cls.ADMISSION_SAMPLES:
        tokens = encode_output(getattr(cls, sample_name))
        accepted = grammar.validate_tokens(tokens)
        admitted = accepted == tokens
        should_admit = sample_name in expected_admission_set
        assert admitted is should_admit, (
            f'Expected structured_outputs admission for {sample_name} '
            f'to be {should_admit}, got {admitted}.'
        )
```

评论区精华

核心讨论是作用域问题：bbrowning 指出 stop_token_ids 被传入所有 grammar 后端，但只有 xgrammar 使用，其他后端是否也会命中同样 bug？yzong-rh 承认 guidance 同样存在问题，即便覆盖 eos_token 也不会掩码 stop tokens，部分后端还需要上游修复，留下大量 TODO。

bbrowning 另外确认本地验证修复了 #51693 的 Inkling 模型 500 错误，并批准合并。

- stop_token_ids 只作用于 xgrammar，其他 grammar 后端是否受影响？ (design): 确认该修复仅覆盖 xgrammar；guidance、llguidance 等后端需单独修复，部分依赖上游能力，以 TODO 跟踪。
- 本地验证是否修复 #51693 的 500 错误 (question): 验证通过，修复对 #51693 场景有效，bbrowning 批准合并。

风险与影响

- 风险：
  1. 移除 <|call|> 依赖 #49227 的 stop token 掩码行为：若 xgrammar 版本回退或后端未启用掩码，Harmony 的 <|call|> 停止行为可能回归。
  2. 修复仅覆盖 xgrammar：guidance、llguidance 等后端仍存在 stop token 未被掩码的问题，用户切换后端可能遇到类似 500 错误或无限生成。
  3. 测试改为 token 级校验并依赖 openai/gpt-oss-20b 的 tokenizer 与 GenerationConfig：离线环境或上游 tokenizer 变化可能导致测试脆弱。 - 影响：用户侧：Harmony 结构化输出在 xgrammar 下的 <|call|> 停止行为恢复正常，Inkling 模型 500 错误被修复。系统侧：HarmonyParser 的 _END_TAG 常量变化影响所有 harmony 结构化请求，测试镜像生产路径后能更早暴露后端回归。团队侧：本次仅清理 xgrammar 路径，其他 grammar 后端留下 TODO，需要后续分别跟进。 - 风险标记：其他 grammar 后端未覆盖，依赖 #49227 的 stop token 掩码，测试依赖远程模型配置

关联脉络

- PR #49227 Mask request stop tokens in xgrammar until grammar terminates: 本 PR 的直接前置：移除 HarmonyParser 补丁依赖 #49227 合并后提供的 stop token 掩码能力，测试也改为通过 XgrammarBackend 传入 stop_token_ids。